SHOUZAB KHAN

SKHAN6

ASSIGNMENT NO 6

"By submitting this homework, I attest that the homework is my own work, and I have not copied any text verbatim from any online or offline source or from other students (past or present). I also attest that I have not used any paraphrasing tool or any homework solution websites. In addition, I attest that any information taken from course slides and handouts or any websites is properly attributed and referenced. I understand that any violation of the UAB Academic Integrity Code for any portion of the homework will result in a zero score for the entire homework and will be reported to the academic integrity coordinator per university policy."

1. In our class discussion on Module 11, we discussed the issues with making large datasets public. We discussed two specific examples: the Massachusetts health insurance dataset, and the Netflix movie recommendation dataset.

Explain the issues that led to privacy violations in each of these two cases. (It is not enough to copy-paste the slide containing mention of the incidents. For full credit, please explain the issues in your own words.) [3+3=6]

**ANSWER:**

Initially, Massachusetts Health Insurance wanted to make the dataset available for use in a different kind of publication. Before the data was to be shared, they (Massachusetts health insurance) deleted important information, including names and addresses. The zip code, gender, and date of birth were the only details provided. They thought it would be impossible to identify someone based just on their zip code, gender, and date of birth. However the privacy researcher found the Massachusetts governor and his medical degree when she compared this data with voter records. She achieved this by comparing voter zip codes, birthdates, and genders to the voting database, which also has these details. She showed how information on a particular person may be found by looking through the two databases.

Secondly, the Netflix film tried to improve the algorithm that suggests moves. In many kinds of apps, the ability to display or suggest user-generated material is considered a positive feature. Netflix expressed alarm over it. Netflix said in the open that a generous reward would be given to the person who creates an effective recommendation algorithm. By making the movie-watching database logs public, they achieved this. Every movie that was stored in Netflix's database was logged. They deleted the real usernames and replaced them with random IDs to maintain secrecy. Someone made use of the knowledge. This database was contrasted with the Internet Movie Database (IMDB). The IMDB database contains information on the film's release date. Users post reviews on this website. It's likely that someone who has been viewing the Avengers movie series has reviewed both the series and other films in similar genres. They can even determine and use the review author's real login from this.

2. In Airavat (the paper discussed in Module 11), how does the system prevent malicious mapper code from opening a network connection to another server (outside the MapReduce servers) and copying the input data? Explain briefly. (2 points)

How is leaking data through extreme mapper output values (e.g., 100000 instead of the range 1-10) prevented by Airavat? (2 points)

**ANSWER:**

Since humans are generally bad code interpreters, examining every piece of code that will interact with your data can occasionally be impossible without knowing what it will accomplish. The answer is to let them run the code without telling them what will happen. The code will execute on data that Airavat permits, but Airavat will retain the data inside the cloud and will stop anyone trying to leak or send the data outside of it. Airavat will check the data output for MapReduce to see whether any information is being leaked. In such a scenario, Airavat will stop the data from being transferred to outside servers.

By these two points, Airavat prevents the huge data output value:

**Mandatory Access Control:**

All data must comply with required access control in order to shield it from outside access. The mapper won't be able to expose the data to the public, even with total access to it. Before uploading the data to the cloud, the data provider will flag the data, making it clear that it is sensitive and shouldn't be shared. The cloud will then enforce the obligatory access control policy, stopping data from being saved through connections, files, and storage channels.

**Differential privacy:**

What happens if the malicious code is written so that, in the event that the user is a, the output is 1? If the user is b, the result will be 2. It is possible to write the code so that it is invisible to the data supplier. This brings us to the idea of differential privacy. From the output, the attacker can infer what input was provided to the data. They look through the data to determine whether anything crucial could leak information about the input; if so, they stop the data from leaving the system to stop leaks from the output. The concept is referred to as functional sensitivity. The amount of information that can be shared will be decided by the data owner. This functional sensitivity will then determine the transfer of information.